CUNY SPS DATA 621 - CTG5 - HW3

Betsy Rosalen, Gabrielle Bartomeo, Jeremy O'Brien, Lidiia Tronina, Rose Koh

April 10th, 2019

Contents

1 DATA EXPLORATION

Relocating to a new city or state can be very stressful. In addition to the stress of packing and moving, you may also be nervous about moving to an unfamiliar area. To better understand their new community, some new residents or people interested in moving to a new city choose to review crime statistics in and around their neighborhood. Crime rate may also influence where people choose to live, raise their families and run their businesses; many potential new residents steer clear of cities with higher than average crime rates.

Data was collected in order to predict whether the neighborhood will be at risk for high crime levels. For each neighborhood the response variable, `target`, represents whetever the crime rate is above the median crime rate or not. In addition to that 13 predictor variables were collected representing each neighborhood's: proportion of large lots, non-retail business acres, whether or not it borders the Charles River, nitrogen oxides concentration, average number of rooms per dwelling, proportion of owner-occupied units, distances to five Boston employment centers, accessibility to radial highways, property tax rate, pupil-teacher ratio, proportion of african Americans, percent lower status, and median value of homes. The evaluation data contains the same 13 predictor variables and no target variable so it will be impossible to check the accuracy of our predictions from the testing data.

VARIABLE NAME	DEFINITION	TYPE
target	whether the crime rate is above the median crime rate (1) or not (0)	response variable
zn	proportion of residential land zoned for large lots (over 25000 square feet)	predictor variable
indus	proportion of non-retail business acres per suburb	predictor variable
chas	a dummy var. for whether the suburb borders the Charles River (1) or not (0)	predictor variable
nox	nitrogen oxides concentration (parts per 10 million)	predictor variable
rm	average number of rooms per dwelling	predictor variable
age	proportion of owner-occupied units built prior to 1940	predictor variable
dis	weighted mean of distances to five Boston employment centers	predictor variable
rad	index of accessibility to radial highways	predictor variable
tax	full-value property-tax rate per \$10,000	predictor variable
ptratio	pupil-teacher ratio by town	predictor variable
black	$1000(B_k - 0.63)^2$ where B_k is the proportion of blacks by town	predictor variable
lstat	lower status of the population (percent)	predictor variable
medv	median value of owner-occupied homes in \$1000s	predictor variable

1.1 Summary Statistics

Looking at the Table 2, we can see that `chas` and `target` are binary variables. 49% of our target variable is coded as 0's indicating that the crime rate is NOT above the median crime rate. There are potential outliers present in `zn`, `lstat`, `medv` and `dis`.

1.2 Shape of Predictor Distributions

Figure. 1 shows that the distribution of most of the variables seems skewed. There are some outliers in the right tail of `tax` , `rad`, `medv`, `lstat`, `dis` and left tail of `ptratio`.

Table 2: Summary statistics

	n	min	mean	median	max	sd
zn	466	0.0000	11.5772532	0.00000	100.0000	23.3646511
indus	466	0.4600	11.1050215	9.69000	27.7400	6.8458549
chas	466	0.0000	0.0708155	0.00000	1.0000	0.2567920
nox	466	0.3890	0.5543105	0.53800	0.8710	0.1166667
rm	466	3.8630	6.2906738	6.21000	8.7800	0.7048513
age	466	2.9000	68.3675966	77.15000	100.0000	28.3213784
dis	466	1.1296	3.7956929	3.19095	12.1265	2.1069496
rad	466	1.0000	9.5300429	5.00000	24.0000	8.6859272
tax	466	187.0000	409.5021459	334.50000	711.0000	167.9000887
ptratio	466	12.6000	18.3984979	18.90000	22.0000	2.1968447
lstat	466	1.7300	12.6314592	11.35000	37.9700	7.1018907
medv	466	5.0000	22.5892704	21.20000	50.0000	9.2396814
target	466	0.0000	0.4914163	0.00000	1.0000	0.5004636

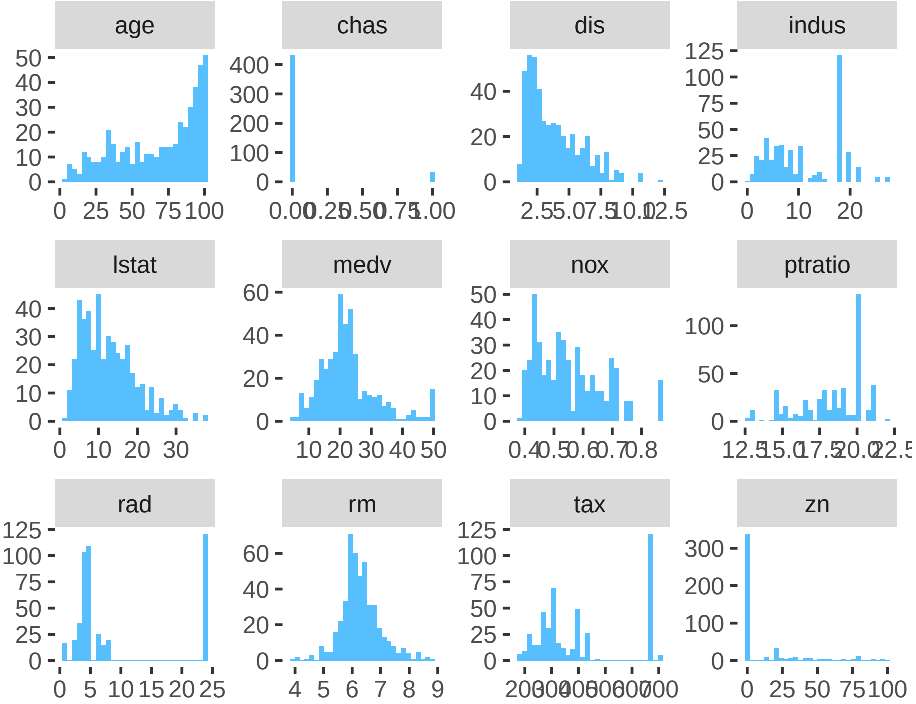

Figure 1: Data Distributions

1.3 Outliers

Figure. 2 shows that there are also a large number of outliers that need to be accounted for, most significantly in `zn` and `medv` and less significantly in `lstat`, `dis` and `rm`. Since `tax` variable has values which are very large compared to other variables in the dataset, it was scaled to fit the boxplot by dividing by 10.

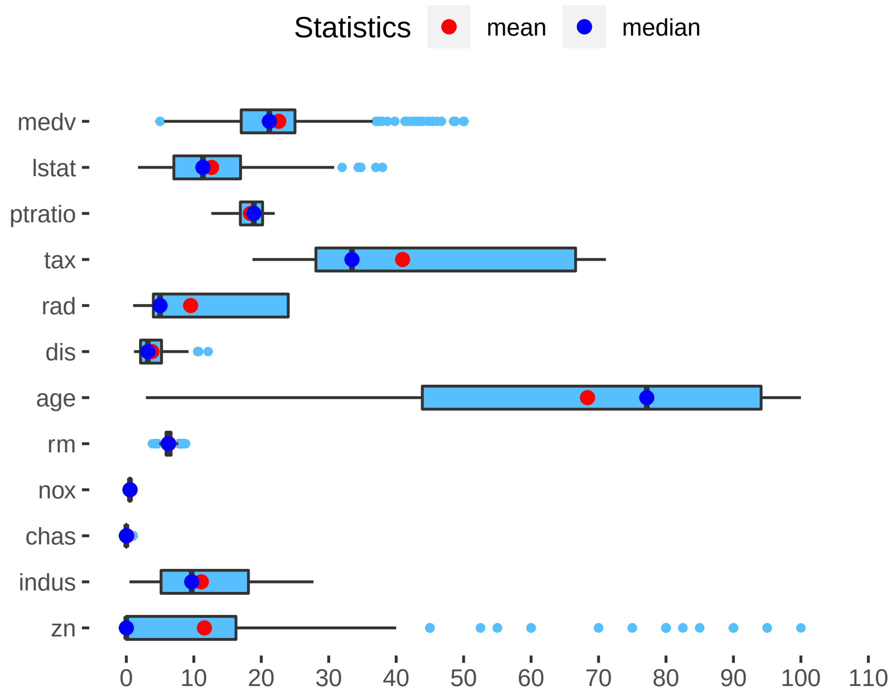

Figure 2: Boxplots highlighting many outliers in the data.

1.4 Missing Values

There are no missing values in any of our observations gathered across the thirteen predictor variables as can be seen in Figure. 3.

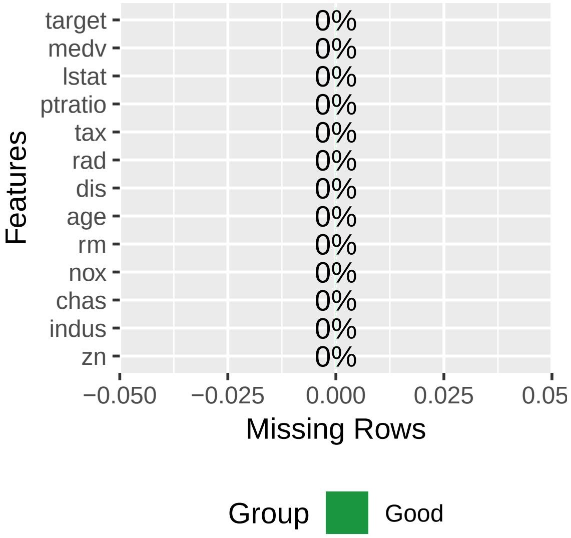

Figure 3: Missing values

1.5 Linearity

Each variable was plotted against the target variable in order to determine at a glance which had the most potential linearity before the dataset was modified.

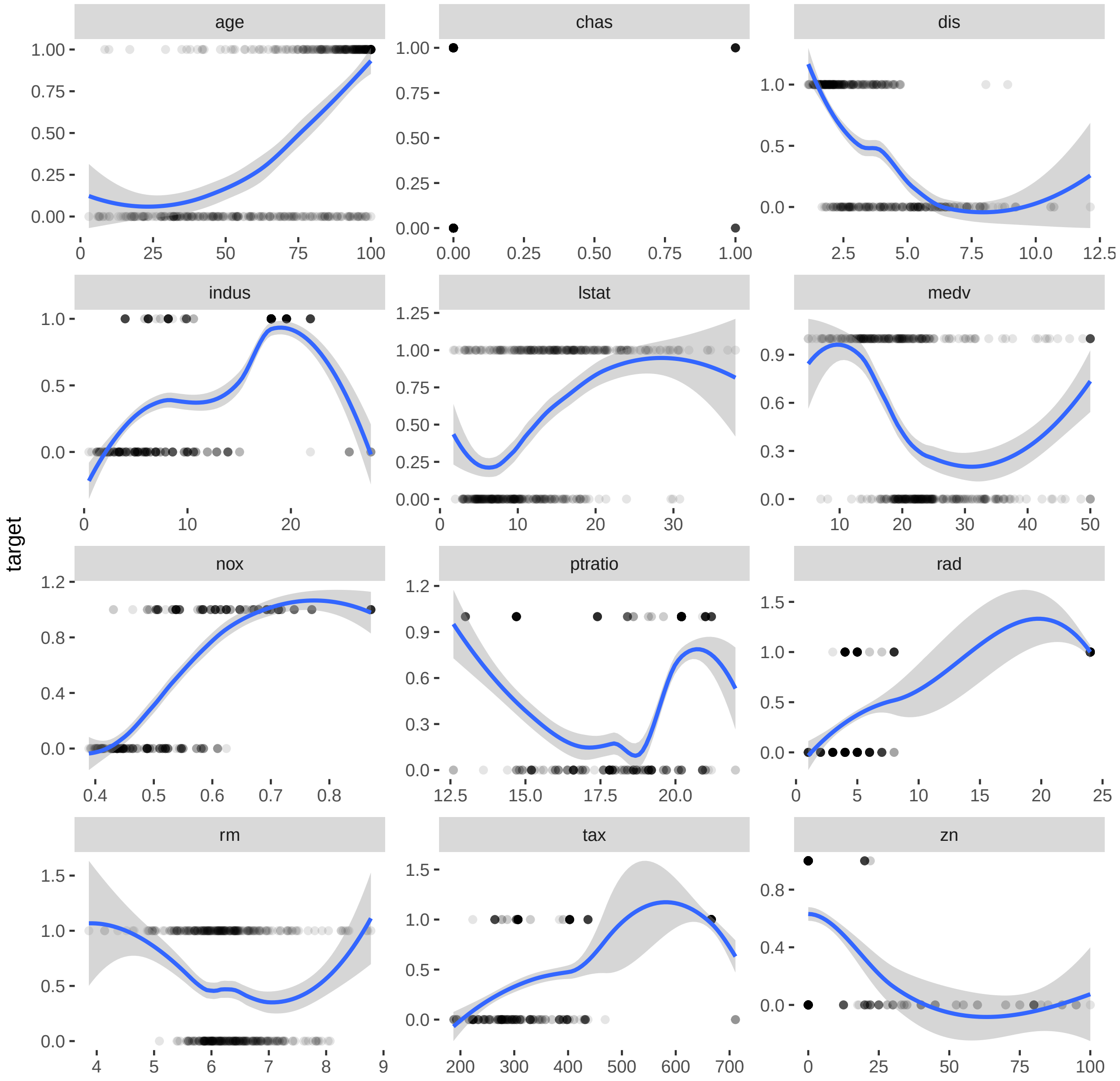

Figure 4: Linear relationships between each predictors and the target

As can be observed in Figure. 4, the most influential variables are the ones previously discussed to have severe outliers and skew, and their linear relationship is negative - the higher the variable, the lower the target wins.

2 DATA PREPARATION

2.1 Missing Values and NA Imputation

Given that the training dataset does include missing values, there's not need to make systematic corrections or imputations.

2.2 Dealing with outliers, leverage, and influence points

While logistic regression can be more robust to leverage points (explantory variable values, which are distant on the x-axis), outliers (response variable values, which are distant on the y-axis) can exert influence which affects the curve and accuracy of target predictions.

- `dis`, `tax` (property tax rate per $10k), and `medv` (median value of owner-occupied homes) see a few outliers and leverage points in both target classes
- `indus` (the non-retail business acreage proportion) and `lstat` (percent lower status population) both have outliers in the below-mean (0) class
- `ptratio` (pupil-teacher ratio) fit is very impacted by density of low values in the above-mean class, making the linear relationship appear parabolic
- `rad` (highway access index) is influenced by a high-value concentration of locations distant from radial highways that fall in the above-mean class
- `rm` (average rooms per dwelling) sees a wider distribution of house size for tha above-mean class then the below-mean; while `zn` (large-lot zoned land proportion) sees the opposite, with a concentration around a few non-residential land proportions for the above-mean class and a wide dispersion for the below-mean class

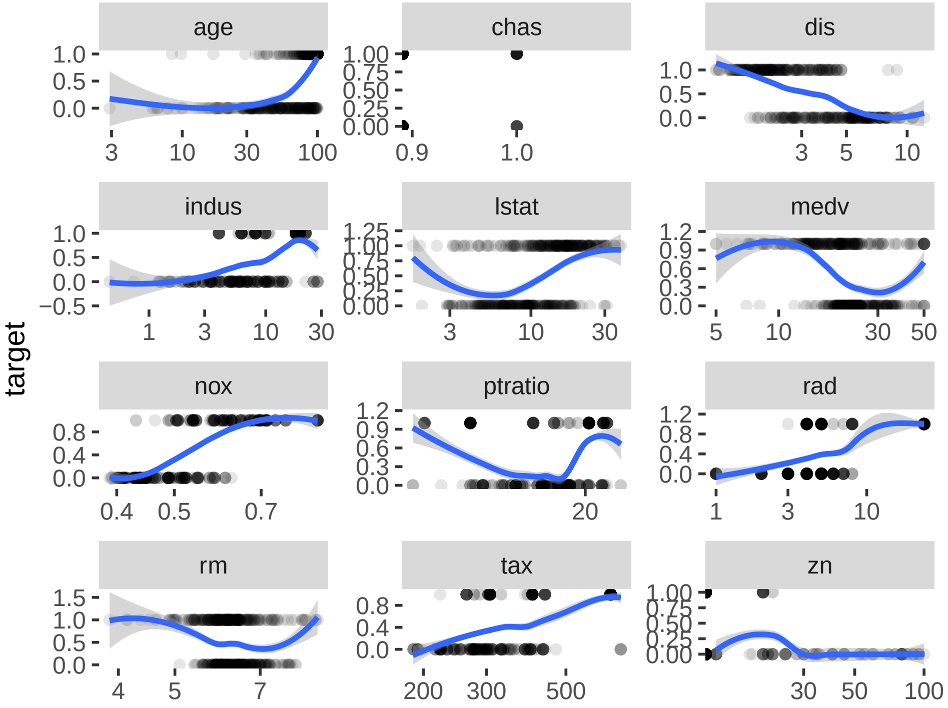

Figure 5: Linear relationships between each predictors and the target

We examined the linear relationships after a log transformation, which smoothed several relationships but still demonstrated visible influence for several variables: `lstat`, `medv`, `ptratio`, `rad`, `rm`, `tax`, and `zn`. We discuss further in the feature engineering section below.

[JEREMY: TEAM, SO WE WANT TO DO FURTHER INVESTIGATION OF OUTLIERS, LOOKING AT R2, STANDARD ERRORS, P-VALS, AND LEVERAGE VALUES FROM THE HAT MATRIX FOR PAIRS OF MODELS, ONE THAT INCLUDES OUTLIERS AND ANOTHER THAT DOESN'T; OR IS THIS ENOUGH?]

2.3 Correlation

```
##                  zn       indus        chas         nox          rm
## zn       1.00000000 -0.53826643 -0.04016203 -0.51704518  0.31981410
## indus   -0.53826643  1.00000000  0.06118317  0.75963008 -0.39271181
## chas    -0.04016203  0.06118317  1.00000000  0.09745577  0.09050979
## nox     -0.51704518  0.75963008  0.09745577  1.00000000 -0.29548972
## rm       0.31981410 -0.39271181  0.09050979 -0.29548972  1.00000000
## age     -0.57258054  0.63958182  0.07888366  0.73512782 -0.23281251
## dis      0.66012434 -0.70361886 -0.09657711 -0.76888404  0.19901584
## rad     -0.31548119  0.60062839 -0.01590037  0.59582984 -0.20844570
## tax     -0.31928408  0.73222922 -0.04676476  0.65387804 -0.29693430
## ptratio -0.39103573  0.39468980 -0.12866058  0.17626871 -0.36034706
## lstat   -0.43299252  0.60711023 -0.05142322  0.59624264 -0.63202445
## medv     0.37671713 -0.49617432  0.16156528 -0.43012267  0.70533679
##                 age         dis         rad         tax    ptratio
## zn      -0.57258054  0.66012434 -0.31548119 -0.31928408 -0.3910357
## indus    0.63958182 -0.70361886  0.60062839  0.73222922  0.3946898
## chas     0.07888366 -0.09657711 -0.01590037 -0.04676476 -0.1286606
## nox      0.73512782 -0.76888404  0.59582984  0.65387804  0.1762687
## rm      -0.23281251  0.19901584 -0.20844570 -0.29693430 -0.3603471
## age      1.00000000 -0.75089759  0.46031430  0.51212452  0.2554479
## dis     -0.75089759  1.00000000 -0.49499193 -0.53425464 -0.2333394
## rad      0.46031430 -0.49499193  1.00000000  0.90646323  0.4714516
## tax      0.51212452 -0.53425464  0.90646323  1.00000000  0.4744223
## ptratio  0.25544785 -0.23333940  0.47145160  0.47442229  1.0000000
## lstat    0.60562001 -0.50752800  0.50310125  0.56418864  0.3773560
## medv    -0.37815605  0.25669476 -0.39766826 -0.49003287 -0.5159153
##               lstat       medv
## zn      -0.43299252  0.3767171
## indus    0.60711023 -0.4961743
## chas    -0.05142322  0.1615653
## nox      0.59624264 -0.4301227
## rm      -0.63202445  0.7053368
## age      0.60562001 -0.3781560
## dis     -0.50752800  0.2566948
## rad      0.50310125 -0.3976683
## tax      0.56418864 -0.4900329
## ptratio  0.37735605 -0.5159153
## lstat    1.00000000 -0.7358008
## medv    -0.73580078  1.0000000
```

An examination of correlation between the explanatory variables reveals the following:

- `indus` (non-retail business acre proportion) is positively correlated with `nox` (pollution concentration, $r = .76$) and `tax` (property tax rate per $10k, $r = .73$) and is negatively correlated with `dis` (weighted mean distance to employment centers, $r = -.7$)
- `chas` (bordering Charles river) correlated with `nox` ($r = .97$) and `rm` (average rooms per dwelling, $r = .91$) and `age` (proportion of pre-1940 homes, $r = .79$); and is negatively correlated with `dis` ($r = -.97$)
- `medv` (median value of owner-occupied homes) is correlated with `rm` ($r = .71$); and is negatively correlated with `lstat` (percent lower status population, $r = -.74$)
- `age` is correlated with `nox` ($r = .74$); and is negatively correlated with `dis` ($r = -.75$)
- `rad` (highway access index) correlated with `tax` ($r = .91$)

[JEREMY: TEAM, LET'S DISCUSS HOW WE'D LIKE TO APPLY THESE CORRELATION FINDINGS TO MODEL EVALUATION AND VARIABLE SELECTION]

2.4 Feature Engineering

In MARR, Sheather quotes Cook and Weisberg, suggesting that the best way to determine need for log transformation of skewed predictors is to include both the original and transformed varaibles in the logistic regression model in order assess their relative contributions directly and prune accordingly

Reexamining the histograms of the predictor distributions above reveals that:

- `age` is left-skewed
- `dis` is right-skewed, and `zn` is extremely so
- `nox` is right-skewed and platykurtic (thin-tailed)
- `rad` and `tax` seem to have normal distributions, with large numbers of outliers at particular levels
- `indus` and `ptratio` reveal pecular skew, with incidences at particular high level, perhaps due to regulation or infrastructure requirements

We include log transforms of `age`, `dis`, `nox`, `rad`, `tax`, `indus`, and `ptratio` in the dataset for evaluation in models.

[JEREMY: TEAM, DO WE WANT TO ADDRESS THIS BY CALLING log() ON VARIABLES WHEN BUILDING glm(), OR SHOULD WE TRANSFORM IN SOURCE DATASET? I'VE ASSUMED THE FORMER.]